[{'title': 'ZERO-SHOT BUILDING AGE CLASSIFICATION FROM FACADE IMAGE USING GPT-4',

'authors': [{'name': 'Z. Zeng',

'affiliation': 'Department of Civil, Environmental and Geomatic Engineering, University College London, UK'},

{'name': 'J. M. Goo',

'affiliation': 'Department of Civil, Environmental and Geomatic Engineering, University College London, UK'},

{'name': 'X. Wang',

'affiliation': 'Department of Civil, Environmental and Geomatic Engineering, University College London, UK'},

{'name': 'B. Chi',

'affiliation': 'Department of Geography, University College London, UK'},

{'name': 'M. Wang',

'affiliation': 'Department of Civil, Environmental and Geomatic Engineering, University College London, UK'},

{'name': 'J. Boehm',

'affiliation': 'Department of Civil, Environmental and Geomatic Engineering, University College London, UK'}],

'abstract': 'A building’s age of construction is crucial for supporting many geospatial applications. Much current research focuses on estimating building age from facade images using deep learning. However, building an accurate deep learning model requires a considerable amount of labelled training data, and the trained models often have geographical constraints. Recently, large pre-trained vision language models (VLMs) such as GPT-4 Vision, which demonstrate significant generalisation capabilities, have emerged as potential training-free tools for dealing with specific vision tasks, but their applicability and reliability for building information remain unexplored. In this study, a zero-shot building age classifier for facade images is developed using prompts that include logical instructions. Taking London as a test case, we introduce a new dataset, FI-London, comprising facade images and building age epochs. Although the training-free classifier achieved a modest accuracy of 39.69%, the mean absolute error of 0.85 decades indicates that the model can predict building age epochs successfully albeit with a small bias.',

'key\_findings': 'The training-free classifier achieved a modest accuracy of 39.69% with a mean absolute error of 0.85 decades, indicating it can predict building age epochs successfully albeit with a small bias.',

'limitation\_of\_sota': 'Building an accurate deep learning model for estimating building age from facade images requires a considerable amount of labelled training data, and the trained models often have geographical constraints.',

'proposed\_solution': 'A zero-shot building age classifier for facade images developed using GPT-4 Vision with prompts that include logical instructions, which does not require any training.',

'paper\_limitations': 'The classifier struggles to predict the age of very old buildings and is challenged by fine-grained predictions within 2 decades.',

'source': 'ZERO-SHOT BUILDING AGE CLASSIFICATION FROM FACADE IMAGE USING GPT-4\nZ. Zeng1, J. M. Goo1, X. Wang1, B. Chi2, M. Wang1, J. Boehm1∗\n1Department of Civil, Environmental and Geomatic Engineering, University College London, Gower Street, London, WC1E 6BT UK\n2Department of Geography, University College London, Gower Street, London, WC1E 6BT UK\n–{zichao.zeng.21, june.goo.21, xinglei.wang.21, bin.chi, meihui.wang.20, j.boehm }@ucl.ac.uk\nKEY WORDS: Building, Facade, Image Understanding, Deep Learning, Multi-modal, Large Vision Language Model\nABSTRACT:\nA building’s age of construction is crucial for supporting many geospatial applications. Much current research focuses on estimating\nbuilding age from facade images using deep learning. However, building an accurate deep learning model requires a considerable\namount of labelled training data, and the trained models often have geographical constraints. Recently, large pre-trained vision\nlanguage models (VLMs) such as GPT-4 Vision, which demonstrate significant generalisation capabilities, have emerged as poten-\ntial training-free tools for dealing with specific vision tasks, but their applicability and reliability for building information remain\nunexplored. In this study, a zero-shot building age classifier for facade images is developed using prompts that include logical\ninstructions. Taking London as a test case, we introduce a new dataset, FI-London, comprising facade images and building age\nepochs. Although the training-free classifier achieved a modest accuracy of 39.69%, the mean absolute error of 0.85 decades indic-\nates that the model can predict building age epochs successfully albeit with a small bias. The ensuing discussion reveals that the\nclassifier struggles to predict the age of very old buildings and is challenged by fine-grained predictions within 2 decades. Overall,\nthe classifier utilising GPT-4 Vision is capable of predicting the rough age epoch of a building from a single facade image without\nany training. The code and dataset are available at https://zichaozeng.github.io/ba classifier.\n1. INTRODUCTION\nIndividual building age is one of the significant attribute data for\nbuilding information modelling and urban digital twins. For the\nbuilding itself, energy demand and housing price are linked to\nbuilding age (Aksoezen et al., 2015; Law et al., 2019). In gen-\neral, building age has considerable impact on historical archi-\ntecture preservation, urban planning and disaster management\n(Sun et al., 2022; Ogawa et al., 2023). However, due to numer-\nous missing data or irregular data collection methods, there are\ninnumerable buildings for which we do not know their age.\nWithin the last decade, many scholars were attempting to estab-\nlish building age using deep learning. Li et al. (2018) presented\na novel method for estimating building age from Google Street\nView images using a convolutional neural network (CNN) for\nimage features extraction and support vector machine for con-\nstruction year regression. Zeppelzauer et al. (2018) introduced\nthe automated age estimation method for unconstrained facade\nphotographs from patch-level visual feature learning to global\nage classification. Afterwards, Despotovic et al. (2019), Sun\net al. (2021), and Ogawa et al. (2023) continually explored pos-\nsible building age estimation methods using deep learning tech-\nniques from facade views.\nWith the emergence of deep learning models, the demand for\ntraining data has grown significantly. Due to the large regional\ndifferences in building architecture, transfer of models between\nregions is uncertain and the number of training images for any\nspecific region is typically limited. Today’s foundation models\nare often trained on data sets with 100s of millions of images,\nsee for example JFT (Hinton et al., 2015) or CLIP (Radford et\nal., 2021). We therefore need to consider for each application\nif training a new model using a new labelled dataset is efforts'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': '',

'limitation\_of\_sota': '',

'proposed\_solution': 'We explore the idea of applying a foundation model without additional training in the context of estimating the age of a building from a single image of its facade.',

'paper\_limitations': '',

'source': 'see for example JFT (Hinton et al., 2015) or CLIP (Radford et\nal., 2021). We therefore need to consider for each application\nif training a new model using a new labelled dataset is efforts\nwell invested, or if using a foundation model performs just as\n∗Corresponding authorgood. We explore this idea in the context of estimating the age\nof a building form a single image of its facade. We do this by\napplying a foundation model without additional training.\n2. RELATED WORK\n2.1 Building Age Epoch\nThe age of building stock in our built environment is a cru-\ncial information for estimating energy demand, urban planning,\ncultural heritage protection, disaster resilience, etc. Building\nage is a crucial indicator in energy consumption analysis and is\nan important variable for energy demand estimation (Aksoezen\net al., 2015; Garbasevschi et al., 2021). During the last few\ndecades, a considerable number of studies on green or energy-\nefficient buildings, have shown that the age of construction is\nhighly related to energy use and has thus become a key indic-\nator to define sustainable construction (Aksoezen et al., 2015).\nThis has also led to building age being closely related to hous-\ning prices, in the real estate industry and in urban planning (Tam\net al., 1999; Stanley et al., 2016; Law et al., 2019). In addition\nto impacts on energy analysis, the assessment of potential dis-\nasters such as earthquake and tsunami damages utilise the year\nof construction as a parameter in damage fragility curve mod-\nels (Nagao et al., 2011; Del Gaudio et al., 2017). Besides, the\nestimation of the building age is essential for the preservation\nof historic buildings and cultural heritage, which can help to\nidentify which buildings need to be preserved or restored, espe-\ncially in urban planning and development.\nDue to the remarkable success of computer vision on partic-\nular applications such as medicine and engineering, images of\nbuildings have been considered as a source for estimate building\nattributes, which can fill gaps in the available data effectively\nand directly. However, most techniques in the study of archi-arXiv:2404.09921v1 [cs.CV] 15 Apr 2024'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': 'In the classification of architectural styles and building age epochs, methods have evolved from using pixelwise classifiers and object detectors to deep convolutional neural networks (CNNs) and Vision Transformers (ViTs). Recent studies have shown that ViTs, such as the Swin transformer, can effectively improve classification accuracy in predicting the age of buildings from street view imagery. Additionally, the integration of Geographic Information System (GIS) data has been explored for enhancing age prediction. The emergence of pre-trained Large Language Models (LLMs) and Vision Language Models (VLMs) has further expanded capabilities in various tasks, including geographic knowledge and reasoning, with zero-shot learning showing promise in fields beyond natural language processing.',

'limitation\_of\_sota': 'Traditional models often struggle when applied to different geographic locations due to the specificity of their training data. This limitation hinders the adaptability and scalability of such models across diverse geographic areas.',

'proposed\_solution': "The article discusses the use of Vision Transformers (ViTs) and the combination of deep CNNs with GIS data for improving the accuracy of building age classification. It also highlights the potential of pre-trained Large Language Models (LLMs) and Vision Language Models (VLMs) in performing tasks across various fields, including geospatial science, without the need for labeled training data. These models' ability to understand both images and language, and their training on large and diverse datasets, allow for better adaptability and application in different geographic locations.",

'paper\_limitations': '',

'source': 'Figure 1. Sample Images contained in FI-London with age epoch\ntectural imagery concentrate on categorising structures accord-\ning to distinct styles of architecture (Despotovic et al., 2019).\nRiemenschneider et al. (2012) developed a method for classi-\nfying architectural styles in images using an irregular rectangu-\nlar lattice and a combination of pixelwise classifiers and object\ndetectors to analyse and segment building facades. Similarly,\nShalunts et al. (2011) categorised different construction styles\nby scale invariant feature transform (SIFT) from facade win-\ndow components. Law et al. (2019) , on the other hand, built\na pipeline based on deep convolutional neural network (CNNs)\nfor extracting visual features from satellite and street view im-\nages and combines them with pre-existing building attributes\n(e.g., age, size, and accessibility) to estimate house prices in\nLondon, UK.\nIn the classification task of building age epochs, Zeppelzauer\net al. (2018) proposed a framework based on CNNs to predict\nbuilding age. The method aggregates patch-level predictions to\nderive a global estimate for the entire building. Despotovic et\nal. (2019) grounded this framework to infer the year of con-\nstruction and then analyse the heating energy demand. Ogawa\net al. (2023) combined geographic information system (GIS)\ndata to predict the age of buildings from street view imagery. In\nthis study, they compared networks based on deep CNNs with a\nViT-based architecture and found that the ViT-based Swin trans-\nformer improved the classification accuracy effectively.\n2.2 Vision Language Model\nPre-trained Large Language Models (LLMs), especially Chat-\nGPT by Achiam et al. (2023), have not only achieved extreme\nperformance in natural language processing (NLP) tasks with\nzero-shot setting (Qin et al., 2023), but have also inspired new\nattempts in other fields. In geospatial science, Roberts et al.\n(2023), Li and Ning (2023) and Wang et al. (2023) indicated\nGPT’s significant capabilities in geographic knowledge and\nreasoning including human mobility prediction, country out-\nlines creation, travel routes planning, supply chain analysis, etc.Notably, these case studies are zero-shot learning examples us-\ning LLMs without any training and tuning, which means no\nlabelled training data is needed to complete prediction tasks in-\ntelligently.\nIn addition to textual tasks, pre-trained models that understand\nboth images and language, often called Vision Language Mod-\nels (VLMs) or multi-modal models, are increasingly used in\nimage understanding tasks. Various large VLMs like CLIP\n(Radford et al., 2021), BLIP (Li et al., 2022a), GLIP (Li et al.,\n2022b), and Grouding-DINO (Liu et al., 2023) have demon-\nstrated their abilities in visual understanding and have attained\na high accuracy with zero-shot learning in visual benchmark\ntasks such as classification, segmentation, detection, and depth\nestimation. These models are trained on a significant number of\nimage-text pairs and are not limited to pre-defined classes. This\nallows large VLMs to identify unlearned object from images by\nan operator’s text input referred to as a prompt (Radford et al.,\n2021).\nBesides being able to identify objects or scenes without prior\nspecific training, VLMs can be used effectively across different\ngeographic locations, thanks to their integration with LLMs that\ninclude training data with large amounts of geographic inform-\nation. Traditionally, models trained in one specific area often\nstruggle when used in a different location (Kedron and Holler,\n2022). However, VLMs that incorporate language understand-\ning can easily adapt and be applied in various geographic areas.\nThis adaptability is due to their training on very large and di-\nverse data, making it easier to bridge geographical differences\n(Zhang et al., 2022).\nFor architectural vision tasks such as building age epoch classi-\nfication, most datasets have restricted geographic locations and'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': 'Large VLMs like GPT-4 emerge as new contenders to classify or predicate structural attributes without any training data.',

'limitation\_of\_sota': 'Most datasets for architectural vision tasks such as building age epoch classification have restricted geographic locations and small sample size.',

'proposed\_solution': 'This study explores the zero-shot ability of GPT-4 as a classical method for classifying or predicating structural attributes without any training data.',

'paper\_limitations': '',

'source': '(Zhang et al., 2022).\nFor architectural vision tasks such as building age epoch classi-\nfication, most datasets have restricted geographic locations and\nsmall sample size. Large VLMs like GPT-4 (Achiam et al.,\n2023) emerge as new contenders to classify or predicate struc-\ntural attributes without any training data. Therefore, in this\nstudy, we explore the zero-shot ability of GPT-4 as a classical'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': '',

'limitation\_of\_sota': '',

'proposed\_solution': '',

'paper\_limitations': '',

'source': '{\n"type": "text",\n"text": """\nYour task is to predict the age epoch of a building in London based on the image provided by users.\nYou will be presented with <building>, an image containing a main building. You need to infer the most likely <\nbuilding\_age\_epoch>.\nOnly select <building\_age\_epoch> from this list: [">2020", "2000-2019", "1980-1999", "1960-1979", "1940-1959",\n"1920-1939", "1900-1919", "1880-1899", "1860-1879", "1840-1859", "1820-1839", "1800-1819", "1750-1799", "1700-1749",\n"<1700"].\nOrganize your answer in the following format containing two keys:\n{\n"age": <building\_age\_epoch>,\n"reason": ""\n}\nThe meaning of two keys:\n- "age": the most likely <building\_age\_epoch> chosen from the provided list.\n- "reason": a concise explanation supporting your prediction. Please do not use line breaks in the reason.\n"""\n}\nFigure 2. The prompt for building age classification used in GPT-4 Vision\nFigure 3. Distribution of Building Epochs in FI-London\nlarge VLM for the building age classification task.\n3. DATA AND EXPERIMENT\n3.1 Dataset\nA related work by Despotovic et al. (2019) created a closed\ndataset that was obtained by web scraping. The dataset contains\na total of 3865 facade images of 2065 buildings. The buildings\nfrom this dataset are all individual houses, which present a more\nhomogeneous architectural style. The chronological classifica-\ntion of this dataset covers modern buildings from 1969 to 2010,\nand this time span makes the chronological differentiation of\nbuilding facade features less obvious. In another related work,\nSun et al. (2021) combined Google Street View data with a\nbuilding age dataset from Amsterdam, aiming for a more de-\ntailed analysis.\nFor the purpose of this study, a dataset of facade images con-\ntaining building-specific attributes of London (referred to as FI-\nLondon) was created. Focusing on Brent and Camden in Lon-\ndon, FI-London contains a total of 131 high-resolution building\nfacade images (6000 ×4000 pixels). These images contain in-\ndividual building facades of varying building types such as res-\nidential apartment blocks, terraced houses, commercial prop-\nerties, etc. Furthermore, the images contain occlusions such as\npedestrians, cars and scaffolding. This is a potential distraction,\nFigure 4. Framework for zero-shot classification of age epoch\nbut a relevant challenge for GPT-4 Vision. FI-London covers 15\ndifferent architectural age epochs seen in Figure 1, derived from\nthe Colouring Cities project by (Hudson et al., 2019; Hudson,\n2023). Due to the unbalanced sample distribution seen in Figure\n3 and small sample size, FI-London can currently only be used\nfor testing. We chose London as a case study location because\nthe city’s buildings vary greatly in age epoch and facade style\n(Jones, 2005). Many of the historic buildings are relatively well\npreserved, providing an ideal test environment for evaluating\nGPT-4 Vision.\n3.2 Building Age Classier\nWe propose a training-free classifier (zero-shot) fto identify\nthe age epochs of buildings using GPT-4 Vision, and the de-\ntailed process is shown in Equation 1 and Figure 4. Firstly, we\nencode the input image xin FI-London Xinto base64 format,\nand the corresponding building age epoch - i.e. ground truth'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': 'The user’s prompt has major influence in LLM-based tasks. Precision and Recall are used to evaluate the model’s performance on individual age epoch, while Micro F1-score is used as a composite metric to evaluate the overall performance on the whole multi-categorisation task. The task of building age estimation is challenging due to the high degree of similarity in architectural styles between consecutive age epochs and the classification into relatively short time intervals. The Mean Absolute Error (MAE) of the age epochs is introduced as another performance metric to quantify the difference between the mid-year of the predicted and actual age epochs.',

'limitation\_of\_sota': '',

'proposed\_solution': 'The study proposes a method for zero-sample classification using GPT-4 Vision for predicting the age epoch of buildings in London. It involves crafting command prompts that include a general task description, the encoded input image, a list of building age epochs to constrain the output, and instructions for formatting the results. The method also requires the output of the reasoning process behind the predictions.',

'paper\_limitations': 'GPT-4 Vision is still unable to directly output JSON-formatted results at the time of the experiments. There is a high degree of similarity in architectural styles between consecutive age epochs, making the task challenging, especially for buildings at the intersection of two epochs.',

'source': '(y′) - is stored as a JSON file for evaluation purposes. We craft\na series of command prompts to perform zero-sample classific-\nation directly using GPT-4 Vision. Finally, the predicted age\nepoch ( y) and descriptive reasoning process ( R) are output by\nthe classifier ( f) with prompts. In the prompts, we revealed to\nGPT-4 that the input image Xis in London. This is because the\ntask of this study is to predict the age epoch, not the specific\ngeographic location.\n(y, R ) =f(x)(x∈X) (1)\nwhere y= Predicted Age Epoch\nR= Descriptive Reasoning Process\nf= GPT-4 Vision with Prompt\nx= Input Image in Test Set X\nThe user’s prompt has major influence in LLM-based tasks.\nSeen in Figure 2, the instruction prompt consists of four com-\nponents. For a start, the instruction prompt includes a general\ntask description for initialising the GPT-4 Vision. Next, the en-\ncoded input image xand the expected age epoch yare defined\nas<building >and<building ageepoch >, respectively, by\nwhich a context is generated to improve GPT-4 Vision’s under-\nstanding of the task and data. Furthermore, since building age\nestimation in this study is regarded as a categorisation task, we\nprovide a list of building age epochs as a way to constrain the\noutput of GPT-4 Vision. Finally, given that GPT-4 Vision is still\nunable to directly output JSON-formatted results at the time of\nour experiments, we have also included instructions for format-\nting the results within the prompt. In addition to the age epochs\ny, we require the output of the reasoning process R, which not\nonly validates the thinking process of GPT-4 Vision, but also\nsupports the subsequent discussion.\n3.3 Evaluation Metrics\nWe employ standard evaluation metrics in the field of image\nclassification to measure the performance of our model. Spe-\ncifically, Precision andRecall are used to evaluate the model’s\nperformance on individual age epoch. Meanwhile, we use Mi-\ncro F1-score as a composite metric to evaluate the overall per-\nformance on the whole multi-categorisation task, whose for-\nmula is given in Equation 2. It is worth noting that Micro F1-\nscore is the same as Accuracy in multi-categorisation task.\nMicro F1 = 2×Precision Micro×Recall Micro\nPrecision Micro +Recall Micro\n=PTPPTP +PFP +PFN\n=PTP\nTotal number of instances=Accuracy(2)\nwhere Micro Precision = the ratio of the total number of\ncorrect predictions across all categories to the total\nnumber of predictions made across all categories.\nMicro Recall = the ratio of the total number of\ncorrect predictions across all categories to the total\nnumber of actual positives across all classes.\nConsidering that there is often a high degree of similarity in\narchitectural styles between consecutive age epochs, and that\nwe classify age epochs into relatively short time intervals of twodecades (5 decades before 1800), the task is rather challenging.\nThis is especially true for buildings that are at the intersection\nof two epochs, where they can easily be misclassified. For this\nreason, we introduced the Mean Absolute Error (MAE) of the\nage epochs as another performance metric (see Equation 3 for\ndetails) to quantify the difference between the mid-year of the\npredicted age epoch myand the mid-year of the true age epoch\nmy′, seen in Equation 3.\nMAE =1\nNNX\ni=1|(myi−my′\ni)/10| (3)\nwhere myi= Mid Year of Predicted Age Epoch\nmy′\ni= Mid Year of Actual Age Epoch\nN= Instance Size\nIn addition, we perform an intuitive analysis of the results by\nconstructing Confusion Matrices . Besides a normal confusion\nmatrix, we design specific confusion matrices in which those\ncases that were incorrectly classified as adjacent ages (i.e., one\nwrong epoch or two wrong epochs) were also considered as cor-\nrect predictions. These specific confusion matrices approach\nallow us to intuitively analyse the degree of chronological pre-\ndiction of GPT-4 Vision.\n3.4 Experiment Setting\nOne of the advantages of our proposed GPT-4-based classifier'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': 'For the 131 building facade images in the FI-London dataset, GPT-4 Vision produced predictions with a total accuracy/micro F1 of 39.69/0.85. The model showed varying degrees of precision, recall, and F1 scores across different age epochs, with the highest precision observed in the 1940-1959 epoch (100%) and the highest recall in the 1960-1979 epoch (75%). The mean absolute error (MAE) varied significantly across epochs, with the lowest MAE observed in the 1980-1999 epoch (0.0 years).',

'limitation\_of\_sota': '',

'proposed\_solution': 'The proposed GPT-4-based classifier does not require any training and the inference process is conducted through the API provided by OpenAI, which does not require significant computational resources.',

'paper\_limitations': 'The paper does not explicitly discuss the limitations of the proposed solution.',

'source': 'allow us to intuitively analyse the degree of chronological pre-\ndiction of GPT-4 Vision.\n3.4 Experiment Setting\nOne of the advantages of our proposed GPT-4-based classifier\nis that it does not require any training and the inference process\nis conducted through the API provided by OpenAI, so it does\nnot require significant computational resources. The experi-\nments were performed on an Ubuntu Server with an Intel(R)\nCore(TM) i5-13600KF @ 3.50GHz, 64GB of random-access\nmemory, and a GeForce RTX 4070 Ti with 12GB of graphic\nmemory. The model we used is GPT-4 Vison Preview . The in-\nference speed was roughly 10 seconds per image, and the cost\nwas $2.08 for 131 images.\n4. RESULT AND DISCUSSION\n4.1 Experimental Result\nFor the 131 building facade images in the FI-London dataset,\nGPT-4 Vision produced predictions, including 52 correct and 79\nAge EpochPrecision Recall F1 MAE\n% % % 10 yrs\n<1700 0.00 0.00 0.00 16.951\n1700-1749 33.33 14.29 20.00 7.212\n1750-1799 42.86 33.33 37.50 2.332\n1800-1819 25.00 62.50 35.71 0.44\n1820-1839 40.00 25.00 30.77 1.06\n1840-1859 0.00 0.00 0.00 2.07\n1860-1879 25.00 16.67 20.00 2.33\n1880-1899 35.29 60.00 44.44 0.40\n1900-1919 25.00 20.00 22.22 2.00\n1920-1939 56.25 56.25 56.25 0.75\n1940-1959 100.00 21.43 35.29 1.57\n1960-1979 28.57 75.00 41.38 0.50\n1980-1999 50.00 40.00 44.44 0.0\n2000-2019 55.56 76.92 64.52 2.00\n>2020 0.00 0.00 0.00 1.001\nTotal Accuracy/Micro F1 39.69 0.85\n1Mid-year my′\niis set to 1700/2020.\n25-decade interval age epochs.\nTable 1. Performance of zero-shot classifier on FI-London'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': "Among the 79 negative cases analyzed, one was identified as a 'hallucination', where GPT-4 generated content outside the given categories but was more accurate. GPT-4 Vision's performance varied significantly across different age epochs, with no correct predictions for epochs <1700, 1840-1859, and >2020, indicating very low performance. However, it showed high performance for epochs 1920-1939 and 2000-2019, with precision, recall, and F1 scores over 50%.",

'limitation\_of\_sota': 'The existing deep learning classifiers may not accurately predict building age epochs due to reliance on training data and the need for optimized instruction cues.',

'proposed\_solution': 'The study suggests that GPT-4 Vision can rely on its wider knowledge to make accurate judgments, indicating a different insight from traditional deep learning classifiers. However, it also highlights the need for further optimization of instruction cues to improve prediction accuracy.',

'paper\_limitations': "The study's limitations include GPT-4 Vision's over-reliance on training data information and the challenge of accurately classifying buildings from less represented or architecturally ambiguous epochs.",

'source': '{\n"age": "1940-1959", # ground truth: 1940-1959\n"reason": "The building’s architectural style suggests a post-war design common between the 1940s and 1950s in London,\ncharacterized by the streamlined, functional appearance and the use of stone cladding.",\n"ID": 55\n} # Correct Result,\n{\n"age": "<1700", # ground truth: 1840-1859\n"reason": "The building in the image has Gothic architectural features, such as pointed arch windows, flying buttresses, and a\nlarge rose window, which are typical of medieval European churches. Given that Gothic architecture was most prevalent\nfrom the 12th to the 16th centuries, the building likely predates the 1700s.",\n"ID": 26\n} # Incorrect Result,\n{\n"age": "1973-1997", # ground truth: 1960-1979\n"reason": "The building in the image appears to be the British Library in London, which was constructed between 1973 and 1997.\nThe architectural style is indicative of the late 20th century, with its large, blocky form, red brickwork, and lack of\nornamentation typical of the Brutalist style which was popular from the 1950s to the mid-1970s but with construction\nperiods extending into the 1980s and 1990s for some Brutalist buildings.",\n"ID": 39\n} # Hallucination Result,\nFigure 5. Output example of a correct result, a incorrect result and a ”hallucination” result\nFigure 6. Facade image (of British Library) of the hallucination\nresult: ground truth y′\ni1960-1979 →age epoch yi1973-1997\nincorrect cases, with our designed no-training classifier. Three\nvisual example results are shown in Figure 5. It is worth not-\ning that among these 79 negative cases, one particular case was\nidentified as a ”hallucination”, i.e., ”the generation of content\nthat strays from factual reality or includes fabricated inform-\nation” (Rawte et al., 2023). In this case, GPT-4 fabricates a\nnew age epoch (1973-1997) which is outside the list of cat-\negories we postulate within the prompt. However, investigating\nthe input image in Figure 6 and analysing the reasoning given\nby GPT-4 in Figure 5, we can find that GPT-4 succeeded in\nrecognising the image as the British Library and gave a more\naccurate prediction. This phenomenon reveals that GPT-4 Vis-\nion can rely to some extent on its wider knowledge to make\naccurate judgements when making predictions, which has dif-\nferent insight from traditional deep learning classifier. Despite\nits more accurate prediction, we defined it as a ”hallucination”\nresult which strays from prescribed categories.\nIn addition, it also suggests that GPT-4 Vision may be over-\nreliant on the information in the training dataset, or that the\ninstruction cues we provide need to be further optimised to im-\nprove prediction accuracy. Overall, this ”hallucination” case\nprovides an interesting perspective on the potential and limita-\ntions of GPT-4 Vision in recognising and understanding build-\ning facade images.\nThrough the subsequent comprehensive analysis of the predic-\ntion results (shown in Table 1), we aim to provide insights into\nthe performance of GPT-4 Vision on building age epoch estim-\nation, assess its accuracy and reliability, and discuss its possible\nimplications for future research on architectural style recogni-\ntion.\nFigure 7. Facade images in age epoch <1700\n4.2 Performance in Each Age Epoch\nThe precision, recall and F1 score of age epochs <1700, 1840-\n1859, and >2020 is 0%, indicating that GPT-4 Vision’s per-\nformance for these age epochs is very low, with no correct pre-\ndictions. This may be due to the insufficient number of building\nfacade images in these age epochs in the testing data, or the ar-\nchitectural features in these age epochs are not obvious.\nGPT-4 Vision, on the other hand, achieves a high performance\nin age epochs 1920-1939 and 2000-2019. The precision, recall\nand F1 score of both are over 50%. Most facade images built\nin these age epochs are classified accurately. However, all other\nepochs are not well classified with low F1 scores (20%-50%).'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': 'Most facade images built in these age epochs are classified accurately with F1 scores over 50%. However, all other epochs are not well classified with low F1 scores (20%-50%). Specifically, the age epoch 1940-1959 has a high precision (100%) but a relatively low recall (21.43%), indicating accurate but not comprehensive prediction capabilities in this age epoch.',

'limitation\_of\_sota': '',

'proposed\_solution': '',

'paper\_limitations': '',

'source': 'and F1 score of both are over 50%. Most facade images built\nin these age epochs are classified accurately. However, all other\nepochs are not well classified with low F1 scores (20%-50%).\nAmong them, age epoch 1940-1959 has a high precision\n(100%) but a relatively low recall (21.43%), which means that\nGPT-4 Vision can perform accurate prediction in this age epoch,'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': 'GPT-4 Vision demonstrates varying levels of precision and recall across different age epochs when classifying building facade images, with certain epochs showing high recall but low precision, indicating a capability to cover most buildings of a specific age but also misclassify many. The Mean Absolute Error (MAE) analysis reveals a generally small bias in predictions, with most age epochs classified into adjacent ones, reflecting the similarity in architectural styles. However, the age epoch 1700-1749 presents a high MAE, suggesting difficulty in prediction due to renovations or rebuilds. The introduction of a special confusion matrix, considering misclassification within one or two neighbouring epochs as correct, significantly improves prediction accuracy for nearly all categories except for the earliest epoch (<1700).',

'limitation\_of\_sota': '',

'proposed\_solution': '',

'paper\_limitations': 'While GPT-4 Vision accurately captures and distinguishes between buildings from widely separated age epochs, it struggles with consecutive epochs that are stylistically similar. Additionally, the model faces challenges in accurately predicting buildings from the earliest age epoch (<1700), even with adjusted confusion matrices.',

'source': 'but ignores a large number of buildings that are constructed\nwithin this age epoch.\nIn contrast, age epoch 1960-1979 has a high recall (75%) but a\nrelatively low precision (28.57%), suggesting that GPT-4 Vis-\nion is able to cover most of the buildings that are actually in this\nage epoch, but at the same time misclassified many samples that\ndo not fall into this age epoch. Similarly, age epoch 1800-1819\nhas 25% precision with 62.5% recall, and age epoch 1880-1899\nhas 35.29% precision with 60% recall.\n4.3 Offset Chronology\nAlthough GPT-4 Vision performed slightly less well in de-\ntailed prediction, by analysing the Mean Absolute Error (MAE)\nwe observed that the bias in the prediction results was not\nlarge with 0.85 decade. For most of the age epochs (1750-\n1799, 1820-1839, 1840-1859, 1860-1879, 1900-1919, 1940-\n1959, 2020-2019, and >2020), the MAEs are only between one\nand three decades, suggesting that building facade images from\nthese chronological periods tend to be classified into adjacent\nage epochs. This phenomenon reflects the similarity of archi-\ntectural styles between these and their adjacent age epochs, as\nwell as in the subtle differences in building facade characterist-\nics and their impact on classification accuracy.\nFor the age epochs 1800-1819, 1880-1899, 1920-1939, 1960-\n1979, and 1980-1999, the MAEs are all less than one decade,\na result that highlights the high predictive accuracy of GPT-4\nVision within these specific age epochs. This not only suggests\nthat the facade features of buildings in these age epochs are rel-\natively distinct, but also implies that these features are similar to\nsome extent to buildings in the most neighbouring age epochs.\nIn particular, as mentioned above, age epochs 1800-1819, 1880-\n1899 and 1960-1979 exhibited higher recall and lower preci-\nsion, while age bands 1920-1939 and 1980-1999 demonstrated\nhigher F1 scores.\nHowever, the prediction difficulty is still high for the age epoch\n1700-1749 with a high MAE of 7.21 decades. The reason may\nbe that these old buildings have been renovated or rebuilt in the\npast time seen in Figure 7, which makes the prediction difficult.\n4.4 General Performance\nThe predictive performance of GPT-4 Vision can be understood\nmore intuitively through the visual analysis of the confusion\nmatrix (shown in Figure 8. In the standard confusion matrix,\nwe observe that GPT-4 Vision’s prediction results show high\nrandomness in certain age epochs, but more prominent perform-\nance in the 1800-1819, 1880-1899, 1920-1939, 1960-1979, and\n2000-2019 age epochs, which have relatively high recall. This\nsuggests that GPT-4 Vision has a better performance in the pre-\ndiction of these specific ages.\nWhen a special confusion matrix is introduced - i.e., misclassi-\nfication of a neighbouring age epoch is also considered as a cor-\nrect prediction - we find that the vast majority of age epochs are\nsignificantly better predicted, with a substantial increase in ac-\ncuracy. The only less significant predictions are for age epochs\n<1700 and 1840-1859.\nFurther, when a special confusion matrix is introduced - i.e.,\nmisclassification of two neighbouring age epochs is also con-\nsidered as a correct prediction - we observe a significant im-\nprovement in prediction accuracy for almost all categories, in-\ncluding the previously underperforming 1840-1859 age epoch.However, prediction accuracy for age epoch <1700 remains a\nchallenge.\nIn general, these findings reveal the capabilities and limita-\ntions of GPT-4 Vision when dealing with the task of classify-\ning building facade images with subtle time-span variations.\nWhile the model is able to accurately capture and distinguish\nage epochs of buildings within a large gaps of years, there is\nstill room for improvement in the predictive accuracy of the\nmodel for consecutive age epochs that are stylistically close to\neach other.\nFigure 8. Confusion Matrices. \*Blue matrix is standard, Orange'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': "The confusion matrices show the model's performance in classifying stylistically close age epochs. The standard blue matrix represents the model's baseline performance. The orange matrix, with adjustment for one adjacent age epoch, and the green matrix, with adjustment for two adjacent age epochs, show improved classification accuracy by considering stylistic similarities between adjacent epochs.",

'limitation\_of\_sota': '',

'proposed\_solution': "The proposed solution involves adjusting the classification model to account for stylistic similarities between consecutive age epochs. This adjustment is quantified in two scenarios: one adjacent age epoch and two adjacent age epochs. These adjustments aim to improve the model's accuracy in distinguishing between stylistically close epochs.",

'paper\_limitations': 'The paper does not specify the limitations of the proposed solution, nor does it provide details on the dataset, the exact improvement metrics, or the potential impact of these adjustments on epochs that are not stylistically close.',

'source': 'model for consecutive age epochs that are stylistically close to\neach other.\nFigure 8. Confusion Matrices. \*Blue matrix is standard, Orange\nmatrix is with adjustment of one adjacent age epoch, and Green\nmatrix is with adjustment of two adjacent age epochs.'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': 'In this study, a training-free classifier based on GPT-4 Vision has been developed to estimate building age epochs from facade images. The classifier achieved a Mean Absolute Error (MAE) of only 0.85 decade in predicting the age epochs of buildings, with 52 correct predictions out of 131 tested images. Most of the age epochs can be predicted successfully but only approximately, indicating they are predicted to be in adjacent but not far off age epochs. However, accurate predictions better than 2 decades and predictions for very old buildings are still challenging.',

'limitation\_of\_sota': '',

'proposed\_solution': 'The study proposed a training-free classifier based on GPT-4 Vision for estimating building age epochs from facade images without the need for training data. An enhanced dataset named FI-London, combining new facade images with stored building attributes, was introduced to implement this solution.',

'paper\_limitations': 'The construction age is difficult to predict particularly for old buildings, possibly because these older buildings have been renovated, which causes the models to be easily confused. Accurate predictions better than 2 decades and predictions for very old buildings are still challenging.',

'source': '5. CONCLUSION\nIn this study, a training-free classifier based on GPT-4 Vis-\nion has been developed to estimate building age epochs from\nfacade images. An enhanced dataset - FI-London combining\nnew facade images with stored building attributes has been pro-\nposed. Since GPT by Achiam et al. (2023) now provides sig-\nnificant generalisation capability which allows to solve some\nparticular tasks without training data, we designed a reasonable\nprompt to implement GPT-4 Vision on building age epoch es-\ntimation. This is an important and specific task in historical\narchitecture preservation, urban planning and disaster manage-\nment. Besides traditional evaluation matrices like precision,\nresult and F1 score, we introduced Mean Absolute Error (MAE)\nto evaluate the year of classifier missed and visualised several\nspecial confusion matrices to show the accuracy in time of the\nclassifier.\nConsequently, in 131 tested images from FI-London, the zero-\nshot classifier predicted 52 correct results and 79 incorrect res-\nults including 1 ”hallucination” result. Based on human cogni-\ntion, this ”hallucination” result was instead very accurate, due\nto the rich a-priori knowledge of GPT-4. Based on the analysis\nof MAE of only 0.85 decade and special confusion matrices, we\nfound that most of the age epochs can be predicted successfully\nbut only approximately, which means they are predicted to be in\nadjacent but not far off age epochs. However, the construction\nage is still difficult to predict particularly for old buildings. The\nreason, based on studies of their input images, may be that these\nolder buildings have been renovated, which causes the models\nto be easily confused. To sum up, GPT-4 Vision has an ability to\nestimate the building age epoch in general, but accurate predic-\ntions better than 2 decades and predictions for very old building\nare still challenging. Furthermore, these results provide valu-\nable insights for further optimisation of the model and highlight\nthe importance of considering architectural style similarities in\nthe building chronology classification task.\nIn the future, we will continue to explore training-free classifi-\ners in other building attributes such as building occupancy. The\nframework is not limited to GPT-4, and other VLMs can be con-\nsidered and compared. Moreover, FI-London will be extended\nto balance the sample size across age epochs.\n6. ACKNOWLEDGEMENT\nZ. Zeng and J. M. Goo are supported by the Engineering\nand Physical Sciences Research Council through an indus-\ntrial CASE studentship with Ordnance Survey (Grant number\nEP/W522077/1 and EP/X524840/1).\nReferences\nAchiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I.,\nAleman, F. L., Almeida, D., Altenschmidt, J., Altman, S.,\nAnadkat, S. et al., 2023. Gpt-4 technical report. arXiv pre-\nprint arXiv:2303.08774.\nAksoezen, M., Daniel, M., Hassler, U. and Kohler, N., 2015.\nBuilding age as an indicator for energy consumption. Energy\nand Buildings 87, pp. 74–86.\nDel Gaudio, C., De Martino, G., Di Ludovico, M., Manfredi,\nG., Prota, A., Ricci, P. and Verderame, G. M., 2017. Empir-\nical fragility curves from damage data on rc buildings afterthe 2009 l’aquila earthquake. Bulletin of Earthquake Engin-\neering 15, pp. 1425–1450.\nDespotovic, M., Koch, D., Leiber, S., D ¨oller, M., Sakeena, M.\nand Zeppelzauer, M., 2019. Prediction and analysis of heat-\ning energy demand for detached houses by computer vision.\nEnergy and Buildings 193, pp. 29–35.\nGarbasevschi, O. M., Estevam Schmiedt, J., Verma, T., Lefter,\nI., Korthals Altes, W. K., Droin, A., Schiricke, B. and Wurm,\nM., 2021. Spatial factors influencing building age predic-\ntion and implications for urban residential energy modelling.\nComputers, Environment and Urban Systems 88, pp. 101637.\nHinton, G., Vinyals, O. and Dean, J., 2015. Distilling the know-\nledge in a neural network. In: NIPS Deep Learning and Rep-\nresentation Learning Workshop.\nHudson, P., 2023. Colouring cities research programme'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': '',

'limitation\_of\_sota': '',

'proposed\_solution': '',

'paper\_limitations': '',

'source': 'ledge in a neural network. In: NIPS Deep Learning and Rep-\nresentation Learning Workshop.\nHudson, P., 2023. Colouring cities research programme\nopen manual, 2023. GitHub: https://github.com/colouring-\ncities/manual/wiki/.\nHudson, P., Dennett, A., Russell, T. and Smith, D., 2019. Col-\nouring london–a crowdsourcing platform for geospatial data\nrelated to london’s building stock. In: Proceedings of the\n27th Annual Gis Research UK Conference, Newcastle Uni-\nversity, Newcastle, UK, pp. 23–26.\nJones, N., 2005. Architecture of England, Scotland, and\nWales. Reference Guides to National Architecture, Blooms-\nbury Academic.\nKedron, P. and Holler, J., 2022. Replication and the search for\nthe laws in the geographic sciences. Annals of GIS 28(1),\npp. 45–56.\nLaw, S., Paige, B. and Russell, C., 2019. Take a look around:\nusing street view and satellite images to estimate house\nprices. ACM Transactions on Intelligent Systems and Tech-\nnology (TIST) 10(5), pp. 1–19.\nLi, J., Li, D., Xiong, C. and Hoi, S., 2022a. Blip: Bootstrapping\nlanguage-image pre-training for unified vision-language un-\nderstanding and generation. In: International Conference on\nMachine Learning, PMLR, pp. 12888–12900.\nLi, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y .,\nWang, L., Yuan, L., Zhang, L., Hwang, J.-N. et al., 2022b.\nGrounded language-image pre-training. In: Proceedings of\nthe IEEE/CVF Conference on Computer Vision and Pattern\nRecognition, pp. 10965–10975.\nLi, Y ., Chen, Y ., Rajabifard, A., Khoshelham, K. and\nAleksandrov, M., 2018. Estimating building age from google\nstreet view images using deep learning (short paper). 10th In-\nternational Conference on Geographic Information Science\n(GIScience 2018) pp. 40:1–40:7.\nLi, Z. and Ning, H., 2023. Autonomous gis: the next-generation\nai-powered gis. arXiv preprint arXiv:2305.06453.\nLiu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C.,\nYang, J., Su, H., Zhu, J. et al., 2023. Grounding dino: Mar-\nrying dino with grounded pre-training for open-set object de-\ntection. arXiv preprint arXiv:2303.05499.\nNagao, T., Yamazaki, F. and Inoguchi, M., 2011. Analysis of\nbuilding damage in kashiwazaki city due to the 2007 niigata-\nken chuetsu-oki earthquake. In: Proc. 32nd Asian Confer-\nence on Remote Sensing, Taipei, Paper, p. 6.'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': '',

'limitation\_of\_sota': '',

'proposed\_solution': '',

'paper\_limitations': '',

'source': 'Ogawa, Y ., Zhao, C., Oki, T., Chen, S. and Sekimoto, Y .,\n2023. Deep learning approach for classifying the built year\nand structure of individual buildings by automatically link-\ning street view images and gis building data. IEEE Journal of\nSelected Topics in Applied Earth Observations and Remote\nSensing 16, pp. 1740–1755.\nQin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M.\nand Yang, D., 2023. Is chatgpt a general-purpose nat-\nural language processing task solver? arXiv preprint\narXiv:2302.06476.\nRadford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G.,\nAgarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J.\net al., 2021. Learning transferable visual models from nat-\nural language supervision. In: International conference on\nmachine learning, PMLR, pp. 8748–8763.\nRawte, V ., Sheth, A. and Das, A., 2023. A survey of\nhallucination in large foundation models. arXiv preprint\narXiv:2309.05922.\nRiemenschneider, H., Krispel, U., Thaller, W., Donoser, M.,\nHavemann, S., Fellner, D. and Bischof, H., 2012. Irregular\nlattices for complex shape grammar facade parsing. In: 2012\nIEEE Conference on Computer Vision and Pattern Recogni-\ntion, pp. 1640–1647.\nRoberts, J., L ¨uddecke, T., Das, S., Han, K. and Albanie, S.,\n2023. Gpt4geo: How a language model sees the world’s geo-\ngraphy. arXiv preprint arXiv:2306.00020.\nShalunts, G., Haxhimusa, Y . and Sablatnig, R., 2011. Ar-\nchitectural style classification of building facade windows.\nIn: International Symposium on Visual Computing, Springer,\npp. 280–289.\nStanley, S., Lyons, R. C. and Lyons, S., 2016. The price effect\nof building energy ratings in the dublin residential market.\nEnergy Efficiency 9, pp. 875–885.\nSun, M., Zhang, F. and Duarte, F., 2021. Automatic build-\ning age prediction from street view images. In: 2021 7th\nIEEE International Conference on Network Intelligence and\nDigital Content (IC-NIDC), IEEE, pp. 102–106.\nSun, M., Zhang, F., Duarte, F. and Ratti, C., 2022. Understand-\ning architecture age and style through deep learning. Cities\n128, pp. 103787.\nTam, C., Tso, T. Y . and Lam, K., 1999. Feng shui and its im-\npacts on land and property developments. Journal of urban\nplanning and development 125(4), pp. 152–163.\nWang, X., Fang, M., Zeng, Z. and Cheng, T., 2023. Where\nwould i go next? large language models as human mobility\npredictors. arXiv preprint arXiv:2308.15197.\nZeppelzauer, M., Despotovic, M., Sakeena, M., Koch, D. and\nD¨oller, M., 2018. Automatic prediction of building age from\nphotographs. In: Proceedings of the 2018 ACM on Interna-\ntional Conference on Multimedia Retrieval, ICMR ’18, As-\nsociation for Computing Machinery, New York, NY , USA,\np. 126–134.\nZhang, Y ., Zhang, F. and Chen, N., 2022. Migratable urban\nstreet scene sensing method based on vision language pre-\ntrained model. International Journal of Applied Earth Obser-\nvation and Geoinformation 113, pp. 102989.'}]